问题解决任务中行动序列的二分类建模：单/两参数行动序列模型*

付颜斌 [1]　陈琦鹏 [1]　詹沛达 [1,2]

([1] 浙江师范大学心理学院，金华，321004)

([2] 浙江省智能教育技术与应用重点实验室，金华，321004)

摘　要　行动序列作为一种典型的过程数据，可反映被试解决问题的详细步骤。鉴于行动序列可区分正误，本文基于二分类 Logistic 建模提出两个复杂度相对较低的行动序列模型——单/两参数行动序列模型(1P-/2P-ASM)；两者差异在于是否允许自由估计问题状态的区分度。通过实证研究和模拟研究对比探究两个新模型与基于多分类 Logistic 建模的序列作答模型(SRM)的表现。研究结果主要发现(1)两个 ASM 能够获得与 SRM 几乎一致的问题解决能力估计值；(2)两个 ASM 的计算耗时明显低于 SRM 的；(3) 2P-ASM 比 1P-ASM 的综合表现更优。总之，两个模型复杂度相对低的 ASM 均能够实现对行动序列的有效分析，有益于行动序列数据分析的落地。

关键词　过程数据，行动序列，问题状态转移，行动序列模型，项目反应理论

---

* 国家自然科学基金青年基金项目(31900795)资助。
通讯作者：詹沛达，E-mail: pdzhan@gmail.com; zhan@zjnu.edu.cn

1 引言

问题解决是指在没有清晰解决方案的任务情境中，个体通过一系列认知加工过程，应用认知技能和认知活动，在问题空间中进行探索，将问题从初始状态转变为问题解决目标状态的过程(Newell & Simon, 1972)。问题解决过程中，被试需要根据问题解决的目标构建计划，选择策略并预估该计划的执行能否达到期望的状态；同时，被试还需要根据问题目标对行动结果进行检查，发现问题并采取补救措施，及时调整先前的行动策略。因此，对问题解决能力的测量，不仅要关注问题解决的最终结果，还需要关注问题解决过程中系列行为(刘耀辉等, 2022)。近年来，国际学生测评项目(Programme for International Student Assessment, PISA) (OECD,2013)推出了模拟生活情境的问题解决测验，通过真实且具有互动性的任务，记录学生在整个问题解决过程中行为的动态变化过程，这为问题解决能力的测量提供了一种全新的方式。类似的，国际成人能力评估项目(PIAAC)通过计算机测量了成人在富技术环境下的问题解决能力。这些测验不仅记录了学生问题解决的结果，还可以将学生在问题解决过程中的操作步骤实时记录在日志文件中，即过程数据(process data)。相较于传统的结果数据，基于过程数据的挖掘分析，可以为推断学生的潜在问题解决能力提供更为丰富的信息。

目前对于过程数据的研究，主要可以分为特征提取与统计建模两类。其中，特征提取可分为理论驱动和数据驱动两种方法，理论驱动的特征提取方法一般采用专家定义的行为指标来对学生的问题解决过程进行评分(Harding et al., 2017; Rosen, 2017; Yuan et al., 2019)，这种方法依赖于专家的知识经验，属于自上而下的特征提取方法。理论驱动方法标定的行为指标不仅能够用作对学生的评分依据，还可以基于一定的测量模型进一步建模分析(Liu et al., 2018; Zhan & Qiao, 2022; Zhang et al., 2022)，但该方法往往要针对不同的任务情境设定不同的特征提取规则，使得应用成本较高。数据驱动的方法指的是应用数据挖掘、机器学习等算法从过程数据中提取信息，常使用的方法包括自然语言处理(He & Von Davier, 2016; Zhan et al., 2015; He et al., 2021)、降维算法(Tang et al., 2016; Tang et al., 2020, Tang et al., 2021)和网络分析方法(Zhu et al., 2016)等。

另外，统计建模大体可分为传统测量模型的迁移应用和随机过程建模两类。传统测量模型的迁移应用主要是先利用特征提取方法提取完成任务的关键指标，然后参照这些关键指标对被试呈现的具体操作或行动序列(action sequence)[1]进行编码(如，若具体操作中包含关键指标则被编码为 1，否则为 0)，最

[1] “行动序列”是指相邻两个问题状态(problem state)之间的状态转移(state transition)；如 A→B 或 AB 表示由当前阶段中的问题状态 A 到下一阶段中问题状态 B 的状态转移；注意，在实践应用中，随着对问题状态的界定不同，存在一个状态转移包含不止一个行动序列的可能；本文中，我们根据语言场景需求交替使用“行动序列”和“状态转移”两个含义相同的名词。

后基于题目作答理论(item response theory, IRT)模型或认知诊断模型对编码数据进行分析，并估计被试的问题解决能力(Liu et al., 2018; Wilson et al., 2017; Zhan & Qiao, 2022; Zhang et al., 2022; Zhu et al., 2016)。然而，这种方法会部分或完全忽视具体操作中的顺序信息。与之相对，已有研究直接对行动序列进行随机过程建模，如动态贝叶斯网络(Levy, 2019)和隐马尔可夫模型(Arieli-Attali et al., 2019; Bergner et al., 2017)。这种方法虽然考虑到了行动序列中的顺序信息，但估计得到的潜变量通常是是离散的属性或知识掌握状态，无法了解被试稳定且连续的问题解决能力。

针对已有方法的局限性，Han 等人(2021)将动态贝叶斯网络与称名作答模型(nominal response model, NRM) (Bock, 1972)相结合，提出了序列作答模型(sequential response model, SRM)。SRM 假设被试的问题解决能力和某行动序列的特征共同决定了被试呈现该行动序列的概率。相比于已有方法，SRM 不仅考虑了行动序列的顺序信息，还可以提供问题解决能力的连续估计值，可用于精细化了解不同被试问题解决能力之间的个体差异。与 NRM 类似，SRM 假设被试在每个问题状态下的所有转移可选项(即行动可选项)都会提供测量信息，进而为任务中每一个可能存在的行动序列都赋予不同的参数(如，转移倾向性参数和转移区分度参数)。本质上讲，SRM 是对行动序列的多分类(或多元无序)建模，即假设下一个阶段中的所有转移可选项之间是没有数量顺序。然而，在实际问题解决任务中，行动序列是有正误之分的：可将有助于成功解决任务的行动序列界定为正确行动序列，而将最终可能会导致任务失败的行动序列界定为错误行动序列。因此，被试在每个问题状态下的所有转移可选项是有正误之分的，并非完全是没有数量顺序的等价关系。

理论上，对于有正误之分的数据，二分类建模更为适宜。与二分类建模相比，多分类建模的相对优势是可以将更丰富的测量信息纳入到数据分析中，但这势必导致模型的复杂性相对更高；更高的模型复杂性通常意味着更多的待估计参数种类和数量，更高的参数估计计算负担，更低的参数估计结果可解释性，更低的实践易用性。基于模型比较与选择的简约原则(Beck, 1943)，本研究拟对包含正误信息的行动序列进行二分类建模，提出单参数和两参数行动序列模型(one- and two-parameter action sequence model, 1P- / 2P-ASM)，以期降低行动序列分析模型的复杂性并增加计算效率；同时，相对简约的模型也有助于增加模型参数估计结果的可解释性，进而增加行动序列模型的实践易用性。

首先，阐述行动序列建模基础并回顾 SRM；其次，介绍本文两个新模型：1P-ASM 和 2P-ASM；然后，基于一则实证研究数据对比两个新模型和 SRM 的参数估计结果，以展现新模型的实践可应用性及其与 SRM 的参数估计结果一致性程度；再然后，通过模拟研究探究两个新模型在不同模拟测验条件的

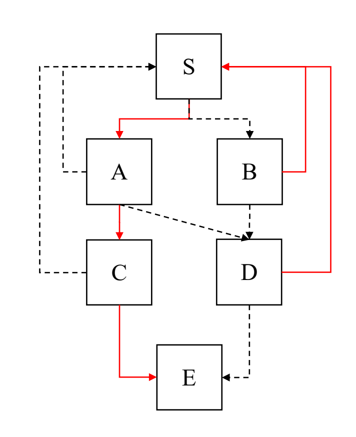
心理计量学性能；最后，对研究结果进行总结并探讨研究局限及未来研究方向。

2 背景知识

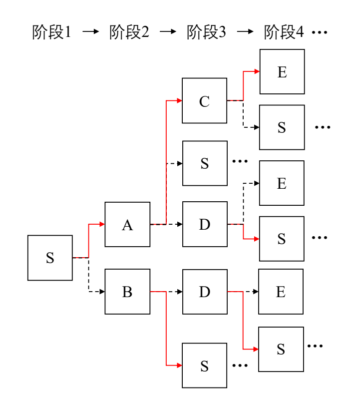
2.1 行动序列建模基础

图 1(a)呈现了一个问题解决任务的例子，该问题解决过程包含了 S、A、B、C、D 和 E 共六种问题状态。其中 S 为问题解决初始状态，E 为问题解决的目标状态，其余均为问题解决的中间状态。由于该题目允许被试在任意中间状态反悔回到初始状态，所以理论上会出现多种问题状态转移序列(problem state transition sequence)，比如，S→A→C→E、S→B→S→A→C→E、S→B→D→E 等。在众多问题状态转移序列中，S→A→C→E 为最优问题状态转移序列，其中包含 S→A、A→C 和 C→E 三个状态转移或行动序列。图中，红色实线箭头表示正确行动序列，即有助于正确解决问题的行动序列；而黑色虚线箭头为错误行动序列，即最终可能导致错误作答结果的行动序列。

实际上，我们可以将被试在每个问题状态下的行动转移视为被试在作答一道"选择题"。图 1(b)是与图 1(a)相对应的问题解决流程图。当被试处于阶段 1 中问题状态 S 时，他/她需要在阶段 2 中的两个问题状态 A 和 B 之间做出选择；同理，当被试处于阶段 2 中问题状态 A 时，他/她需要在阶段 3 中三个问题状态 C、D 和 S 之间做出选择(S 表示返回到初始状态)。此时，我们就可将适用于题目层面作答精度数据分析的传统 IRT 模型迁移应用于此。比如，Han 等人(2021)就将 NRM 迁移应用于此，进而基于多分类建模提出了 SRM。




(a) 各问题状态转移关系图



(b) 问题解决流程图

图 1 问题解决任务示意图

注：红色实线箭头表示正确行动序列，黑色虚线箭头表示错误行动序列；S→A→C→E 为最优问题状态转移序列，其中包含 S→A、A→C 和 C→E 三个行动序列。省略号表示问题解决流程的重复出现。

2.2 SRM 简介

假设一个问题解决任务包含了 R 种离散的问题状态，问题状态的集合为$\boldsymbol{x} = \{x_1, x_2, \ldots, x_R\}$；$S_{n,p} \in \boldsymbol{x}$ 表示学生n在阶段p $(p = 1, \ldots, P_n)$上所处的问题状态，其中P_n为学生n最终呈现问题状态转移序列的长度，不同学生的问题状态转移序列的长度不尽相同。图 2(a)呈现了 SRM 的逻辑示意图，即 SRM 假设被试的问题解决能力影响其在相邻两阶段之间的状态转移；图 2(b)呈现了 SRM 的建模示意图，即 SRM 实际上是对状态转移或行动序列进行建模，假设被试的问题解决能力影响被试呈现特定行动序列的概率。SRM 可表示为：

$$P\left(Y_{n(S_p \to S_{p+1})} = x_j \to x_k \middle| \theta_n\right) = P\left(S_{n,p+1} = x_k \mid S_{n,p} = x_j,\ \theta_n\right) = \frac{exp(\lambda_{x_j x_k} + I_{x_j x_k}\theta_n)}{\sum_{x_h \in \mathbf{M}_{p+1}} \exp\left(\lambda_{x_j x_h} + I_{x_j x_k}\theta_n\right)}\ , \quad (1)$$

式中，$Y_{n(S_p \to S_{p+1})}$为观察变量，即被试 n 在相邻阶段间呈现的行动序列$x_j \to x_k$；$x_j \in \mathbf{M}_p$ 和 $x_k \in \mathbf{M}_{p+1}$分别表示当前阶段所处的问题状态和下一阶段可以选择的问题状态，$\mathbf{M}_p \in \boldsymbol{x} = \{x_1, x_2, \ldots, x_R\}$ 表示在阶段 p 所有可能出现状态集合。θ_n 为被试 n 的问题解决能力；$\lambda_{x_j x_k}$为状态转移倾向参数，表示从状态x_j向状态x_k转移的倾向性，该参数值越大表明行动序列$x_j \to x_k$越易于被呈现；$I_{x_j x_k}$为状态转移区分度参数，该参数值越大表明行动序列$x_j \to x_k$对问题解决能力的区分度越高。SRM 假设给定被试潜在能力后各相邻阶段呈现的行动序列之间满足条件独立，进而，被试最终呈现的行动序列向量$\boldsymbol{Y}_n = (S_{n1} \to S_{n2}, \ldots, S_{np} \to S_{n,p+1})$'的联合概率为：

$$\mathrm{P}(\boldsymbol{Y}_n \mid \theta_n) = \prod_{p=1}^{p_n - 1} P\left(S_{n,p+1} \middle| S_{n,p},\ \theta_n\right)\ . \quad (2)$$

作为一种多分类模型，SRM 中的每一个行动序列都包含 2 个参数，$\lambda_{x_j x_k}$和$I_{x_j x_k}$。仍以图 1(b)为例，SRM 将每一阶段的“选择题”视为“称名作答题”，认为每一个选项都会提供测量信息，进而包含了 22 个参数，分别为 11 个转移倾向性参数(如，λ_{SA}、λ_{SB}、λ_{AS}、λ_{AC}、λ_{BD}和λ_{DE})和与之对应的 11 个转移区分度参数。为了使模型可识别并降低待估计参数数量，Han 等(2021)对 SRM 进行了一定约束：(1) 约束当前问题状态x_j与下一阶段中所有可选的问题状态的之间的转移倾向参数合为 0，即$\sum_{x_k \in \mathbf{M}_{p+1}} \lambda_{x_j x_k} = 0$；(2)预先固定转移区分度参数：若$x_j \to x_k$为正确行动序列，则$I_{x_j x_k} = 1$；若$x_j \to x_k$为错误行动序列，则$I_{x_j x_k} = -1$。

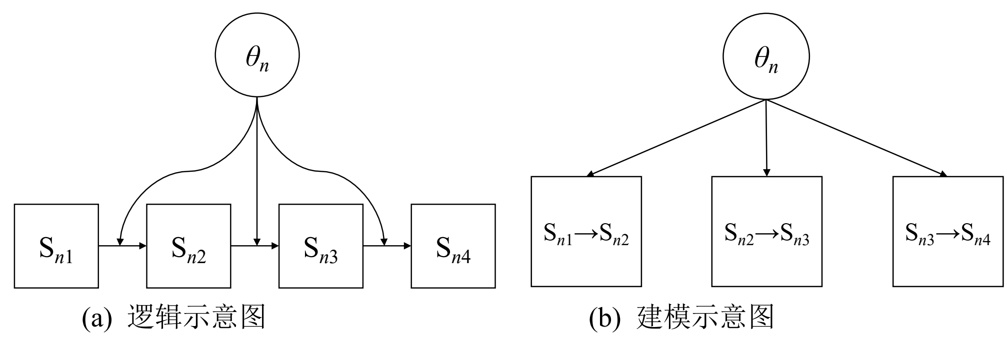


图 2 序列作答模型示意图

注：θ_n 为被试 n 的问题解决能力；S_{n1}为学生n在阶段 1 所处的问题状态，依此类推；$S_{n1} \rightarrow S_{n2}$为学生$n$从阶段 1 向阶段 2 转移的行动序列，依此类推.

3 两个约束模型：1P-ASM 和 2P-ASM

3.1 模型构建

尽管 SRM 采用多分类建模将所有行动序列所提供的测量信息均纳入到模型之中，但它仍然通过一个预先设定的状态转移区分度参数区别对待了行动序列的正确与否。针对具有正误之分的行动序列，本研究采用二分类建模思路，使用针对二级评分数据的 IRT 模型对行动序列进行建模，如单参数 IRT 模型/罗氏模型(Rasch, 1960)和两参数 IRT 模型(Birnbaum, 1968)。对此，图 3 呈现了与图 1 对应的问题解决任务的二分编码示意图，该图中我们将正确行动序列编码为 1，错误行动序列编码为 0。图 3(b)中，我们可以将每一阶段中的“选择题”视为“具有正确答案的多项选择题”；此时，就可以借鉴针传统二级评分 IRT 模型来构建行动序列模型。

图 4 呈现了两个 ASM 的建模示意图。首先，将任务中所有的行动序列进行二分编码：将正确行动序列编码为 1，将错误行动序列编码为 0。此时，被试解决问题所呈现的行动序列向量就被编码为仅包含 0 或 1 元素的二元向量；比如图 1 中最优问题状态转移序列 S→A→C→E 所对应的行动序列向量(SA, AC, CE)’可被转换为(1,1,1)’。然后，基于二级评分 IRT 模型，假设被试的问题解决能力影响被试呈现正确行动序列的概率。

借鉴单参数 IRT 模型，1P-ASM 可被表示为：

$$P\left(Y_{n(S_p\rightarrow S_{p+1})}=1\middle|\theta_n\right)=P\left(Y_{n,x_j\rightarrow}=1\middle|\theta_n\right)=\frac{\exp\left(\beta_{x_j\rightarrow}+\theta_n\right)}{1+\exp\left(\beta_{x_j\rightarrow}+\theta_n\right)}\ , \tag{3}$$

式中，$Y_{n(S_p\rightarrow S_{p+1})}=1$表示被试 n 在相邻阶段间呈现了正确行动序列；$\beta_{x_j\rightarrow}$为行动容易度(action easiness)参数，表示状态x_j下呈现正确行动序列的容易度；其他参数含义同上。

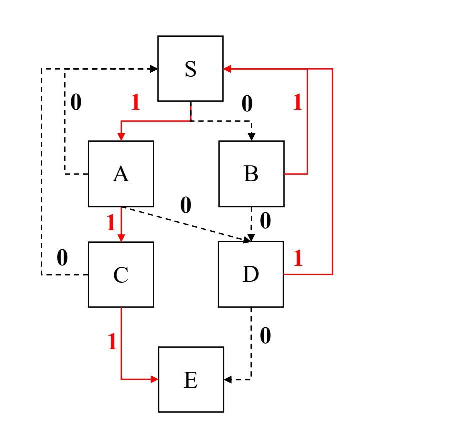
借鉴两参数 IRT 模型，2P-ASM 可被表示为：

$$P\left(Y_{n(S_p\to S_{p+1})}=1\middle|\theta_n\right)=P\left(Y_{n,x_j\to}=1\middle|\theta_n\right)=\frac{\exp\left(\beta_{x_j\to}+\gamma_{x_j\to}\theta_n\right)}{1+\exp\left(\beta_{x_j\to}+\gamma_{x_j\to}\theta_n\right)}\ ,\tag{4}$$

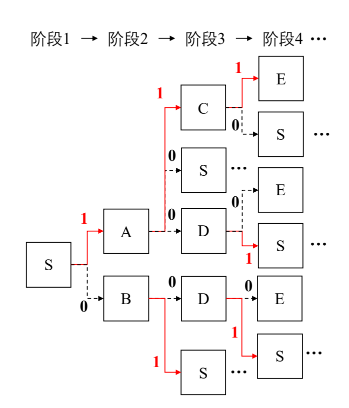
式中，$\gamma_{x_j\to}$ 为行动区分度(action discrimination)参数，表示状态 x_j 下呈现正确行动序列对问题解决能力的区分程度；其他参数含义同上。

与 SRM 类似，两个 ASM 也假设给定被试潜在能力后各相邻阶段呈现的行动序列之间满足条件独立，进而，被试最终呈现的行动序列二元向量$\boldsymbol{Y}_n$的联合概率为：

$$\mathrm{P}(\boldsymbol{Y}_n\ |\theta_n)=\prod_{p=1}^{p_n-1}P\left(Y_{n(S_p\to S_{p+1})}\middle|\theta_n\right)\ .\tag{5}$$



(a) 各问题状态转移关系图



(b) 问题解决流程图

图 3 问题解决任务二分编码示意图

注：红色实线箭头表示正确行动序列，编码为 1；黑色虚线箭头表示错误行动序列，编码为 0；省略号表示问题解决流程的重复出现。

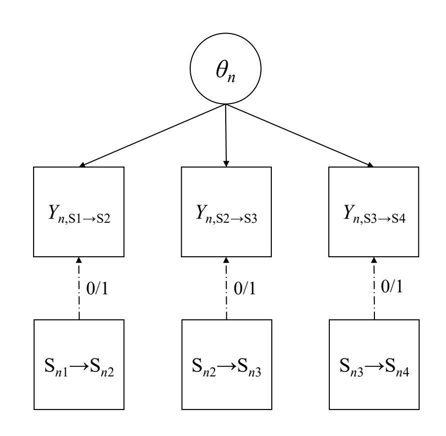


图 4 二分类行动序列模型建模示意图

注：θ_n 为被试 n 的问题解决能力；S_{n1}为学生n在阶段 1 所处的问题状态，依此类推；$S_{n1}\to S_{n2}$为学生n从阶段 1 向阶段 2 转移的行动序列，依此类推；$Y_{n,\mathrm{S1}\to\mathrm{S2}}$ 为二分编码后的行动序列，$Y_{n,\mathrm{S1}\to\mathrm{S2}}=1$ 表示被试 n 呈现了正确行动序列，$Y_{n,\mathrm{S1}\to\mathrm{S2}}=0$ 表示被试 n 呈现了错误行动序列。

3.2 ASM 和 SRM 的对比

ASM 和 SRM 最大的区别在于建模逻辑不同，前者是二元 logistic 模型，后者是采用除总模型形式的多分类 logistic 模型。建模逻辑上的差异不仅会导致模型复杂性的差异，也会导致参数解释上的差异。比如，如果将 SRM 中的转移倾向性参数视为“选项”层面的参数，那两个 ASM 中的行动容易度参数就是“题目”层面的参数；前者刻画选择某选项的倾向性(即，呈现某行动序列的倾向性)，而后者刻画答对该题目的容易度(即，呈现正确行动序列的容易度)。另外，为了减少参数估计数量，SRM 中的状态转移区分度为预先固定的，无需参数估计；而 2P-ASM 中的行动区分度参数为自由估计参数，可以反映不同问题状态(或“题目”)对被试问题解决能力的区分程度。

值得注意的是，由于 SRM 中的参数约束，其待估计参数的数量并不总是多于 1P-ASM 和 2P-ASM。以如图 1(b)中阶段 3 的问题状态 C 为例，当下一阶段的转移可选项只有 2 个时(E 和 S)，由于 SRM 约束了$\lambda_{CE}+\lambda_{CS}=0$、$I_{CE}=1$和$I_{CS}=-1$，所以 SRM 中也仅需估计 1 个转移倾向性参数。此时，1P-ASM 也仅需估计 1 个行动容易度参数，而 2P-ASM 还需要额外估计 1 个行动区分度参数。当然，SRM 的待估计参数数量会随着下一阶段的转移可选项的增加而增加，而 ASM 则不会。

3.3 贝叶斯参数估计

与 SRM 一样，两个 ASM 也可使用全贝叶斯马尔可夫链蒙特卡洛(MCMC)算法进行参数估计。本研究使用 R 软件中的 Rstan 包(Guo et al., 2018)完成 MCMC 参数估计，Rstan 默认使用 No-U-Turn Sampler (NUTS) (Hoffman & Gelman, 2014)作为抽样方法。网络版附录 5 中呈现了两模型参数估计对无信息先验分布和有信息先验分布的鲁棒性分析结果，结果表明无论先验分布包含的信息量如何，两模型的参数估计结果均具有较高鲁棒性。结合已有研究结果(Han et al., 2021; Fu et al., 2022)，正文中所有参数估计均采用有信息先验分布。参数估计代码及示例数据已经分享在 https://osf.io/3y2xr/?view_only=7bc05393a51f472aa2462214ba588063

4 实证数据分析

4.1 任务描述

与 Han 等人(2021)研究保持一致，本研究也选用 PISA 2012 计算机化问题解决“Tickets”任务(CP038Q02)的行动序列数据进行分析。该任务要求被试操作一台虚拟售票机，购买一张可以乘坐 2 次

的全价郊区火车票。图 5 呈现了该任务的初始界面，问题解决过程中各阶段的截图见附录图 A1。为解决问题，被试首先需要在交通方式上选择“城市地铁”或“郊区火车”。其次，根据所选的交通方式，被试需要在“全价票”和“打折票”之间做选择。然后，根据所选票价类型，再选择购买“包日票”或“次票”；如果选择“次票”则还要选择购买的乘车次数(“1 次”~“5 次”)。最后做出“购买”决定即可完成该任务。被试可以在任意操作界面通过点击“取消”来返回到任务的初始界面重新进行选择。为了解决该任务，不同被试最终呈现的问题状态转移序列(或与其对应的行动序列向量)的长度不尽相同。

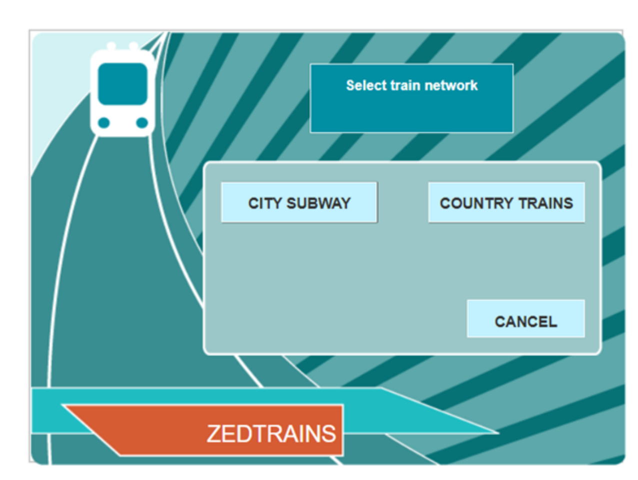


图 5　PISA 2012 购票任务初始界面

图 6 呈现的是该任务拆解后的问题结构，共包含 11 个问题状态，即$\boldsymbol{x} = \{S, A, B, C, D, E, F, G, H, I, J\}$；其中 S 为起始问题状态，J 为终止问题状态，其余均为中间问题状态。在两个相邻问题状态间，实线表示正确状态转移(如，SA)，虚线表示错误状态转移(如，SF)。该任务的最优问题状态转移序列为“开始(S)→正确的交通类型(A)→正确的折扣类型(B)→正确的车票类型(C)→正确的乘车次数(D)→购买(J)”，相应的点击操作是“乡村火车”→“全价票” →“次票” →“2 次” →“购买”。

表 1 从“选择题”视角进一步整理了图 6 中的操作过程。可将当前阶段所处的问题状态视为一道被试需要作答的“选择题”，将下一阶段的可选问题状态视为“选项”。比如，在初始阶段被试需要在“选择题”S 的两个“选项”A 和 F 之间进行选择；其中 A 为正确“选项”，F 为错误“选项”。针对这些“选择题”，SRM 将它们视为称名作答题，ASM 将它们视为二级评分选择题。比如，某学生的问题状态转移序列为 SABCDEDJ，则 SRM 分析的行动序列向量为(SA, AB, BC, CD, DE, ED, DJ)’，而 ASM 分析的行动序列二分向量为(1, 1, 1, 1, 0, 1, 1)'。

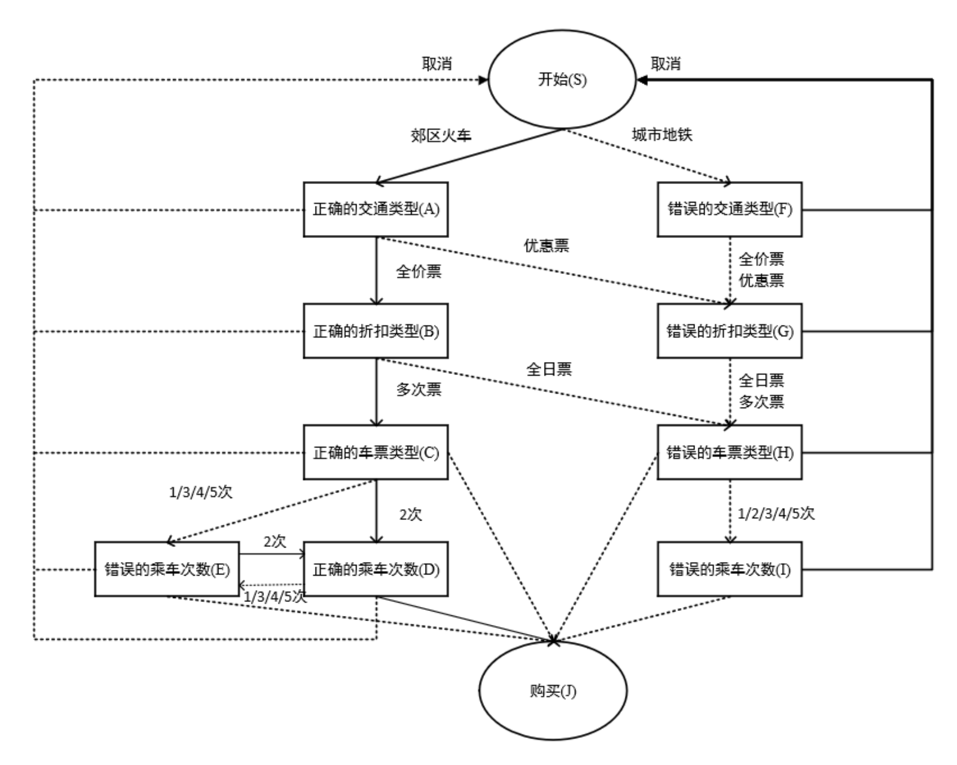


图 6 PISA 2012 购票任务结构图

表 1 PISA 2012 购票任务所类比的“选择题”.

当前问题状态	下一阶段可选问题状态(转移选项)			
S	A (1)	F (0)		
A	B (1)	G (0)	S (0)	
B	C (1)	H (0)	S (0)	
C	D (1)	E (0)	S (0)	J (0)
D	J (1)	E (0)	S (0)	
E	D (1)	J (0)	S (0)	
F	S (1)	G (0)		
G	S (1)	H (0)		
H	S (1)	I (0)	J (0)	
I	S (1)	J (0)		

注：括号中的 1 代表正确“选项”(即正确状态转移)，0 代表错误“选项”(即错误状态转移)。

4.2 数据整理与分析

原始数据来源于 PISA 官网下载[2]。在进行具体的数据分析之前，先根据图 6 中定义的任务结构对原始数据进行重新编码，并对数据进行清理： (1) 删去提前终止作答的行动序列，即没有点击“购买”的行动序列；(2) 删除包含了不可能的状态转移的行动序列(如附录表 A1)。最终，从记录行动的日志文件中

[2] https://www.oecd.org/pisa/pisaproducts/database-cbapisa2012.htm

提取了 28851 名被试的问题状态转移序列(行动序列向量)，其中问题状态转移序列的最短长度为 5，最长长度为 110，平均长度为 6.992。原始数据当中包含了 1395 种问题状态转移序列，其中有 569 种问题状态转移序列完成了任务目标(涉及 15408 名被试：有 10610 名被试按照最优问题状态转移序列完成了任务目标，另外 4798 名学生在正确解决问题过程中有错误修正过程)。最后，限于算力且为增加研究效率，我们从 28851 名被试中随机选择了 2000 名学生的问题状态转移序列用于本研究的实证分析(问题状态转移序列的最短长度为 5，最长长度为 46，平均长度为 7.03；包含了 1395 种问题状态转移序列，其中有 569 种问题状态转移序列完成任务目标[涉及 1068 名学生，有 737 人按照最优问题状态转移序列完成了任务目标)。

分别使用 1P-ASM、2P-ASM 和 SRM 分析数据。参数估计时，选用 2 条马尔可夫链，每条链长 5,000 次，预热(burn-in)3,000 次。使用 PSRF 值 (PSRF; Gelman & Rubin, 1992) 来确定 MCMC 算法得到的参数估计值是否达到收敛；当 PSRF < 1.1 时，表明参数估计收敛。此外，采用 Watanabe-Akaike 信息准则 (WAIC; Watanabe & Opper, 2010) 和留一法交叉验证 (LOO, Gelfand et al., 1992)两个完全贝叶斯的相对拟合指标来衡量模型对数据的拟合情况，为模型选择提供证据；两个指标值越小，表明模型对数据的拟合越好。值得注意的是，由于 SRM 和 ASM 分析的数据并不相同(前者分析的是每位学生的行动序列向量，后者分析的是每位学生的行动序列向量的二分化向量)，所以两者的相对拟合值无法比较。因此，我们仅能通过相对拟合指标判断两个 ASM 之间的相对拟合优劣，无法用于判断 ASM 和 SRM 的相对拟合优劣。对此，本研究将通过计算 ASM 和 SRM 参数估计结果的一致性来体现二分类建模具有与多分类建模相接近的表现。另外，使用后验预测检验(PPC; Gelman et al., 2014) 评估模型对数据的绝对拟合；如果模型拟合数据，则其后验预测概率(*ppp*)接近 0.5，反之，如果模型不拟合数据，则其 *ppp* 值 < 0.025 或 > 0.975。本文中 PPC 所使用的统计量见附录表 A2。

4.3 结果

所有模型中所有参数的 PSRF 值均小于 1.05，表明在我们的设定下所有参数估计达到收敛标准。此外，附录 4 中提供了模型参数的抽样轨迹图。表 2 呈现了三个模型对数据的拟合情况和计算耗时。首先，三个模型的 *ppp* 值均接近 0.5，表明三个模型均拟合该数据。其次，两个相对拟合指标表明 2P-ASM 对数据的拟合优于 1P-ASM，意味着考虑行动序列的区分度能更好地反映该数据的特征，即不同行动序列对问题解决能力的区分能力是不同的。如上文所述，ASM 和 SRM 的相对拟合结果不具有可比性。最

后，参数估计耗时可以综合反映模型的复杂性程度，结果发现 SRM 的耗时最长，2P-ASM 次之，1P-ASM 的耗时最短；这表明二分类模型的确比多分类模型简约。下文主要研究结果围绕两个 ASM 阐述，并呈现 ASM 和 SRM 对被试问题解决能力估计的一致性。

表 2　实证研究中三个模型对数据的拟合情况和计算耗时.

模型	LOO	WAIC	*ppp*	计算时间(秒)
1P-ASM	11018.208	11007.133	0.511	647.5
2P-ASM	10363.785	10275.475	0.518	958.5
SRM	16804.501	16803.925	0.498	1958.6

注: 1P-ASM = 单参数行动序列模型；2P-ASM = 两参数行动序列模型；SRM = 序列作答模型; LOO = 留一法交叉验证；WAIC = Watanabe-Akaike 信息准则；*ppp* = 后验预测概率。

表 3 中呈现了两个 ASM 的题目参数估计结果(后验均值、后验标准差和 95%最高概率密度[贝叶斯可信区间])。首先，对于行动容易度参数而言，正确问题解决路径(即最优问题状态转移序列)上的问题状态(S、A、B、C 和 D)的容易度参数的后验均值均大于 0，表明当被试处于正确路径上的问题状态时，其更容易继续呈现正确行动序列；与之相对，错误问题解决路径(即非最优问题状态转移序列)上的问题状态(F、G、H 和 I)的容易度参数后的验均值均小于 0，表明当被试已经处于错误路径上的问题状态时，其更难以纠正错误转向正确的问题状态(即更易于继续维持在错误路径上)。值得注意的是，问题状态 E 和 I 是错误路径上的问题状态，其含义均为“选择错误的乘车次数”；相较于其他错误路径上的问题状态，E 和 I 的容易度估计值更高，表明当被试处于这两个错误状态时，更有可能在下一步选择时纠正自己的错误(即选择 S 返回初始状态重新作答)。其次，对于行动区分度参数而言，不同问题状态的行动区分度有一定差异性。其中，问题状态 C 和 I 的行动区分度后验均值相对较高，表明不同问题解决能力的学生在这两个问题状态下呈现正确行动序列的概率差异相对较大。也就是说，已处于正确问题解决路径上的学生是否能够选择正确的乘车次数，以及已经处于错误问题解决路径上的学生是否能够通过“取消”来纠正自己的错误，这两个操作对于学生的能力的区分力是相对最强的。总之，根据行动参数估计值可发现，当被试已经处于正确问题解决路径，则其更易于保持在正确问题解决路径上；而当被试已经处于错误问题解决路径，则其更易于继续错下去，直到末尾选择乘车次数界面时才有一个纠正错误的关键期。

表 3　实证研究中行动序列模型参数估计结果.

当前问题状态	1P-ASM			2P-ASM					
	容易度			容易度			区分度		
	后验均值	后验标准差	95% HPD	后验均值	后验标准差	95% HPD	后验均值	后验标准差	95% HPD
S	0.911	0.046	(0.822, 1.001)	0.969	0.057	(0.860, 1.084)	1.343	0.116	(1.111, 1.570)
A	1.553	0.066	(1.425, 1.682)	1.547	0.077	(1.401, 1.701)	1.457	0.212	(1.043, 1.870)
B	1.432	0.073	(1.290, 1.577)	1.354	0.082	(1.198, 1.521)	1.797	0.398	(0.958, 2.566)
C	1.436	0.080	(1.279, 1.599)	1.207	0.148	(0.940, 1.526)	3.015	0.885	(1.104, 4.759)
E	2.008	0.107	(1.801, 2.215)	1.734	0.159	(1.456, 2.099)	1.615	0.495	(0.463, 2.576)
D	0.361	0.176	(0.015, 0.702)	0.472	0.283	(-0.031, 1.064)	1.472	0.952	(0.225, 3.792)
F	-1.705	0.107	(-1.918, -1.492)	-1.590	0.123	(-1.829, -1.348)	1.438	0.250	(0.982, 1.974)
G	-1.888	0.111	(-2.105, -1.677)	-1.747	0.147	(-2.050, -1.480)	2.115	0.354	(1.495, 2.875)
H	-0.749	0.075	(-0.898, -0.599)	-0.292	0.172	(-0.636, 0.037)	2.229	0.426	(1.400, 3.088)
I	-0.368	0.157	(-0.686, 0.062)	0.760	0.470	(-0.101, 1.753)	3.127	0.933	(1.525, 5.179)

注: 1P-ASM＝ 单参数行动序列模型；2P-ASM＝ 两参数行动序列模型；SRM＝ 序列作答模型；95% HPD＝95%最高概率密度(贝叶斯可信区间).

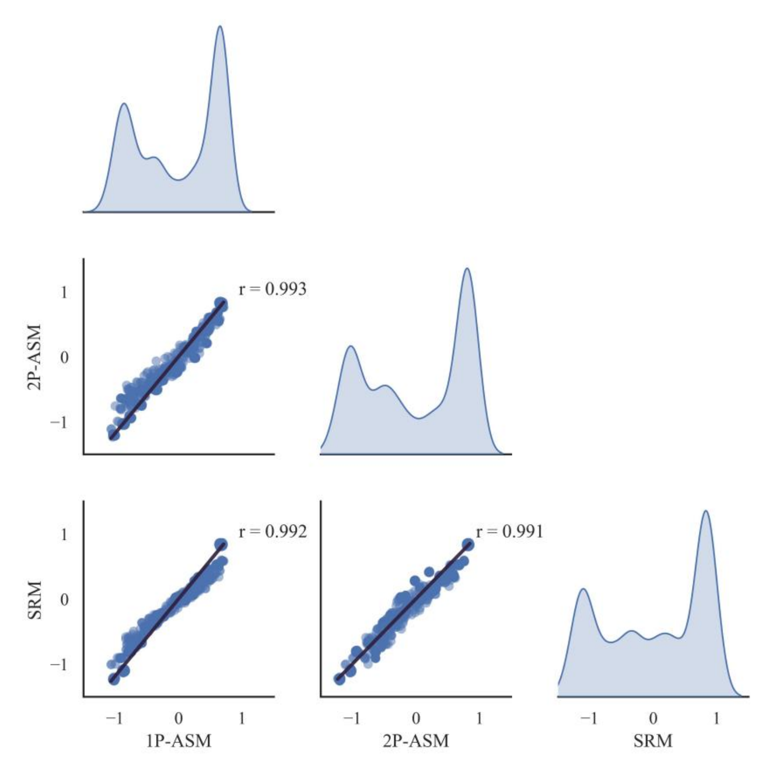


图 7 实证数据中三个模型的问题解决能力参数后验均值对比散点图及概率密度图

注: 1P-ASM=单参数行动序列模型；2P-ASM=两参数行动序列模型；SRM=序列作答模型；r=皮尔逊积差相关。

图 7 呈现了三个模型的问题解决能力估计值(后验均值)的对比散点图及概率密度图。首先，散点图结果呈现出三个模型的问题解决能力估计值具有较高的一致性(三者之间的相关系数均在 0.99 以上)，表明它们测量的是同一潜在特质且二分类建模与多分类建模一样能够通过分析行动序列数据测量被试的问题解决能力并反映个体之间的差异性。其次，对比三模型的概率密度图，可发现三个模型在高能力区间和低能力区间的概率密度分布基本一致，仅在中能力区间的分布略有差异(主要是 SRM)。一个可能的原因是 SRM 更充分地利用了不同行动序列所提供的测量信息：它不仅利用了正确行动序列所包含的测量信息，也利用了不同错误行动序列中可能并不相同的测量信息。比如，当多名被试同时处于问题状态 A 时，相比于选择错误“选项”G 的被试而言，选择错误“选项”S 的被试的问题解决能力似乎要更高一些；此时，SRM 是可以区分呈现 AG 的被试和呈现 AS 的被试之间的区别的，而 ASM 则将他们均视为同一类做出错误选择的人。

从分析数据中挑选取出现频率大于 20 次的问题状态转移序列作为典型问题状态转移序列(涵盖了 80.1%的被试)。表 4 呈现了典型问题状态转移序列在三个模型中的问题解决能力估计值的描述统计(按

SRM 的能力估计均值从高到低排序)。首先，三个模型对呈现各典型问题状态转移序列的被试的能力估计的描述性统计具有一定的一致性。比如，呈现最优问题状态转移序列 SABCDJ 的被试的能力估计均值相对最高，而呈现最差问题状态转移序列 SFGHIJ 的被试的能力估计均值相对最低。其次，整体而言，各典型问题状态转移序列中，出现正确问题状态的数量越多且出现错误问题状态的数量越少则被试的能力估计值的均值就越高，反之，被试的能力估计值的均值就越低。然后，对比 ASM 和 SRM 的结果，发现 ASM 中有两个序列下的被试的能力估计值的均值排序与 SRM 中的不同：SABCEDJ 对应的能力估计值的均值略低于 SFGHSABCDJ 对应的。呈现 SABCEDJ 的被试尽在状态 C 上的选择出现了错误转移(CE)且马上进行了纠正(ED)，而呈现 SFGHSABCDJ 的被试在初始状态就出现了错误转移，直到选择购买乘车次数时才返回初始页面纠正自己的错误。ASM 和 SRM 在这两个序列上的排序差异可以从不同的视角解释。首先，从出现错误状态的次数或问题解决效率(序列长度)看，似乎呈现 SABCEDJ 的被试的能力估计值均值应该高于呈现 SFGHSABCDJ 的被试的；SRM 的排序结果支持该视角解释。其次，结合表 3 中的行动容易度参数可发现，问题状态 C 的容易度较高(难度较低)，而问题状态 F、G 和 H 的容易度较低(难度较高)；因此，从错误选择对能力估计带来的负面影响或惩罚看，在状态 C 的错误选择所带来的惩罚高于在状态 F、G 和 H 的错误选择所带来的，进而导致 SABCEDJ 的被试的能力估计值均值低于呈现 SFGHSABCDJ 的被试的；ASM 的排序结果支持该视角解释。

最后，鉴于相对拟合指标无法对比 ASM 和 SRM 对数据的拟合优劣，我们使用三个模型的问题解决能力估计值对该任务的作答精度数据(根据该任务的评分规则，购买到正确的车票得 1 分，否则得 0 分)做 logistic 回归。按照 logistic 回归的要求，对自变量即能力估计值进行了标准化处理。计算得到的 SRM、1P-ASM 和 2P-ASM 的能力估计值的回归系数分别是 15.195、14.762 和 15.201 (回归系数均显著 $p < 0.001$)，表明三模型的能力估计值的变化可以显著影响该任务的成果完成与否，且影响程度基本一致，其中 2P-ASM 的影响相对最大，SRM 的次之，1P-ASM 的相对最小。

表 4 典型问题状态转移序列对应的问题解决能力估计值的描述统计.

问题状态转移序列	频数	SRM			1P-ASM			2P-ASM		
		均值	中位数	标准差	均值	中位数	标准差	均值	中位数	标准差
SABCDJ	737	0.837	0.837	0.011	0.821	0.821	0.009	0.666	0.665	0.012
SFSABCDJ	35	0.525	0.525	0.007	0.676	0.677	0.007	0.604	0.603	0.01
SFGSABCDJ	22	0.345	0.347	0.007	0.598	0.598	0.005	0.488	0.487	0.009
SABCEDJ	23	0.279	0.279	0.007	-0.017	-0.018	0.005	0.257	0.258	0.008
SFGHSABCDJ	52	0.152	0.151	0.006	0.304	0.304	0.004	0.295	0.296	0.009
SABCJ	47	0.023	0.025	0.007	-0.238	-0.237	0.006	-0.035	-0.035	0.011
SABCEJ	27	-0.25	-0.25	0.005	-0.404	-0.404	0.005	-0.338	-0.338	0.012
SABHJ	117	-0.364	-0.364	0.007	-0.506	-0.506	0.006	-0.359	-0.358	0.01
SAGHJ	65	-0.662	-0.662	0.008	-0.741	-0.742	0.008	-0.594	-0.594	0.01
SAGHIJ	45	-0.806	-0.806	0.008	-0.94	-0.939	0.007	-0.76	-0.76	0.01
SFGHJ	337	-1.099	-1.099	0.011	-1.033	-1.033	0.008	-0.869	-0.869	0.011
SFGHIJ	95	-1.228	-1.227	0.011	-1.201	-1.201	0.008	-1.02	-1.021	0.01

注：1P-ASM = 单参数行动序列模型；2P-ASM = 两参数行动序列模型；SRM = 序列作答模型。

5 模拟研究

5.1 研究设计、数据生成与分析

通过一则模拟研究进一步探究两个 ASM 在理想测验情境下的心理计量学表现。需要强调的是 ASM 本身并无法生成被试解决任务所呈现的问题状态转移序列(只能生成二分向量)；因此，模拟研究中使用 SRM 作为行动序列数据的生成模型。采用实证研究中的问题解决任务结构(图 6)来生成行动序列数据。模拟研究包含两个操纵变量：样本量(含 100、200 和 500 人三个水平)和问题状态转移序列长度(含短和长两个水平)；参照 Han 等人(2021)和 Fu 等人(2022)的做法，在 SRM 中通过调整“取消”操作(如，A→S)的转移倾向参数来操作问题状态转移序列的长度：参数取值越大问题状态转移序列长度越长。网络版附录中表 A3 呈现了短序列和长序列中所有状态转移倾向参数的“真值”。本研究中生成的短问题状态转移序列和长问题状态转移序列的平均长度分别约为 10.5 和 20.2。被试的问题解决能力参数依标准正态分布随机生成。为减少随机误差影响，每种模拟条件下重复生成 50 组数据，每组数据中被试的问题解决能力参数重新生成，但状态转移参数固定不变。

针对生成的数据，使用 SRM、1P-ASM 和 2P-ASM 进行参数估计，参数估计过程与实证研究中保持一致；同样使用 PSRF 作为参数估计收敛指标。由于 SRM 和两个 ASM 建模逻辑不同，它们除了问题解决能力参数含义相同且可比，其余参数含义不同且无法比较。对此，我们针对问题解决能力的估计结果从两方面来评估模型的表现。首先，从参数估计精度方面考虑，使用 Bias 和均方根误差(RMSE)来探究三个模型中问题解决能力参数的估计返真性：$Bias(\theta)=\frac{\sum_{r=1}^{50}\widehat{\theta_r}-\theta_r}{50}$，$RMSE(\theta)=\sqrt{\frac{\sum_{r=1}^{50}(\widehat{\theta_r}-\theta_r)^2}{50}}$，式中$\theta_r$ and $\widehat{\theta_r}$ 分别表示在第 r 次重复中的能力参数的“真值”和参数估计值；此外，还计算了“真值”和估计值之间的相关系数 Cor。其次，从参数估计一致性方面考虑，使用一致性偏差(CBias)和一致性误差(CRMSE)来探究两个 ASM 的能力估计值和数据生成模型 SRM 的能力估计值之间的一致性：$CBias(\hat{\theta})=\frac{\sum_{r=1}^{50}\widehat{\theta_{SRM}}-\widehat{\theta_{ASM}}}{50}$，$CRMSE(\hat{\theta})=\sqrt{\frac{\sum_{r=1}^{50}(\widehat{\theta_{SRM}}-\widehat{\theta_{ASM}})^2}{50}}$，其中，$\widehat{\theta_{SRM}}$表示第 r 次重复中 SRM 的能力估计值，$\widehat{\theta_{ASM}}$表示第 r 次重复中 ASM 的能力估计值；此外，还计算了两类模型估计值之间的相关系数 CCor。

另外，计算了各个条件下 50 次参数估计平均计算时间(ART)来评估不同的模型的参数估计效率以反映模型的复杂性：$ART=\frac{\sum_{r=1}^{50}T_r}{50}$，式中，$T_r$ 表示第 r 次重复中的参数估计计算时间。为保证计算时间结果可比，三模型的所有程序均在相同服务器上运行(配置为 Intel(R) Xeon(R) Gold 6266C CPU @

3.00GHz 和 64G 内存)。

5.2 结果

首先，在所有条件下，三模型中所有参数的 PSRF 均小于 1.1，表示所有模型参数估计均收敛。表 5 呈现了不同模拟条件下三个模型的问题解决能力参数估计的返真性和计算耗时。首先，样本量对能力参数估计的返真性的影响较小；序列平均长度越长，能力参数估计的返真性越高。其次，SRM 作为数据生成模型，其返真性理应最好，2P-ASM 次之，1P-ASM 最差，但三者间整体差异不大(绝大多数条件下 1P-ASM 的 RMSE 比 SRM 的高不到 0.05，Cor 低不到 0.02)。最后，在所有条件下 1P-ASM 的计算耗时最短，2P-ASM 次之，SRM 最长；该结果与实证研究结果吻合，表明相比于多分类模型，二分类建模在保证其能力参数估计精度仅有微弱下降的同时，可大幅减少参数估计耗时。

表 6 呈现了不同模拟条件下两个 ASM 与 SRM 的问题解决能力参数估计的一致性。整体看，两个 ASM 与 SRM 的一致性均较高，且 2P-ASM 与 SRM 的一致性高于 1P-ASM 与 SRM 的一致性。另外，值得注意的是，当序列长度增加后，1P-ASM 与 SRM 的一致性略有下降，而 2P-ASM 与 SRM 的一致性略有提升。

表 5　模拟研究中三个模型的问题解决能力参数的估计返真性和计算耗时.

样本量	序列长度	模型	均 Bias	均 RMSE	Cor	ART(秒)
100	短	1P-ASM	-0.002	0.534	0.854	18.117
		2P-ASM	-0.002	0.534	0.852	30.274
		SRM	0.007	0.515	0.863	1029.189
	长	1P-ASM	-0.011	0.441	0.910	24.393
		2P-ASM	-0.011	0.408	0.917	37.100
		SRM	-0.026	0.395	0.921	1321.361
200	短	1P-ASM	0.007	0.523	0.855	41.923
		2P-ASM	0.007	0.518	0.858	66.395
		SRM	0.011	0.507	0.864	527.740
	长	1P-ASM	0.010	0.438	0.912	54.448
		2P-ASM	0.010	0.395	0.921	76.707
		SRM	0.002	0.386	0.924	691.308
500	短	1P-ASM	-0.004	0.516	0.856	119.439
		2P-ASM	-0.004	0.504	0.863	198.838
		SRM	-0.001	0.500	0.865	590.051
	长	1P-ASM	0.005	0.444	0.907	160.661
		2P-ASM	0.005	0.394	0.920	236.195
		SRM	0.002	0.391	0.921	801.767

注: 1P-ASM＝ 单参数行动序列模型；2P-ASM＝ 两参数行动序列模型；SRM＝ 序列作答模型；均 Bias ＝ 所有被试的估计偏差的均值；均 RMSE ＝ 所有被试的均方根误差的均值；Cor ＝ 真值与估计值之间的相关系数；ART＝ 平均计算时间。当样本量为 100 时，SRM 模型的计算耗时明显多于其他较高样本量条件下的计算耗时；可能是因为样本量较少的情况下，数据提供的测量信息有限，使复杂程度较高的 SRM 的 MCMC 抽样更为困难.

表 6 模拟研究中两个 ASM 和 SRM 的问题解决能力参数估计的一致性.

样本量	序列长度	1P-ASM 与 SRM			2P-ASM 与 SRM		
		均 CBias	均 CRMSE	CCor	均 CBias	均 CRMSE	CCor
100	短	-0.009	0.126	0.991	-0.009	0.129	0.989
	长	0.015	0.200	0.986	0.015	0.098	0.995
200	短	-0.004	0.126	0.991	-0.004	0.098	0.994
	长	0.008	0.208	0.986	0.008	0.077	0.997
500	短	-0.002	0.126	0.990	-0.002	0.060	0.998
	长	0.003	0.211	0.986	0.003	0.042	0.999

注: 1P-ASM = 单参数行动序列模型；2P-ASM = 两参数行动序列模型；SRM = 序列作答模型；均 CBias = 所有被试的一致性偏差的均值；均 CRMSE = 所有被试的一致性误差的均值；CCor = SRM 估计值与 ASM 估计值之间的相关系数.

6 总结与讨论

与传统作答精度数据相比，诸如行动序列等过程数据能提供有关被试如何解决问题的更丰富信息。同时，行动序列数据的非标准化格式(即不同被试的数据长度不同)也给传统心理计量学模型的直接应用带来了困难。近些年，随着研究者们的推进，针对行动序列的分析方法或模型逐渐丰富，主要包括基于特征(关键行动/操作)提取的传统心理计量模型迁移应用和随机过程建模。针对已有方法或多或少的局限，Han 等人(2021)将动态贝叶斯网络与 NRM 相结合，提出了 SRM。与 NRM 类似，SRM 采用多分类 logistic 建模，进而为任务中每一个可能存在的行动序列都赋予不同的参数，导致模型复杂性较高。鉴于问题解决任务中行动序列有正误之分，而非是没有数量顺序的等价关系，本文基于二分类建模提出了两个模型复杂性相对较低的行动序列模型——1P-ASM 和 2P-ASM。不同与 SRM 将 NRM 迁移应用至行动序列数据分析，1P-ASM 和 2P-ASM 分别将更为简单的单参数 IRT 模型和两参数 IRT 模型迁移应用至行动序列数据分析。实证研究结果发现(1)两个 ASM 和 SRM 的问题解决能力估计值具有接近于 1 的相关系数，表明它们测量的是同一潜在特质；(2) 两个 ASM 的计算耗时明显低于 SRM 的，表明 ASM 的模型复杂性低于 SRM 的；(3)当被试已经处于正确问题解决路径，则其更易于保持在正确问题解决路径上；反之，当被试已经处于错误问题解决路径，则其更易于继续错下去；(4)与 1P-ASM 和 SRM 将区分度参数进行固定不同，2P-ASM 可以提供在当前所处问题状态下呈现正确行动序列的区分度参数，有助于确定相对比较重要的问题状态(比如实证研究中的问题状态 C 和 I)，以便数据分析者更好地了解任务本身。模拟研究结果发现(1)即便不是数据生成模型，两个 ASM 也能提供较高的参数估计返真性；(2)

两个 ASM 的计算耗时低于 SRM，尤其是在小样本量条件下的相对优势更为明显；(3)两个 ASM 的问题解决能力估计值与 SRM 的均具有很高的一致性，且 2P-ASM 与 SRM 的一致性相对更高；(4)被试解决问题时最终呈现的问题状态转移序列的长短是影响两个 ASM 以及 SRM 参数估计返真性的主要原因之一：序列越长，数据所含信息越多，对问题解决能力的估计精度更高。综上所述，本文基于二分类建模提出的两个 ASM 能够实现对行动序列数据的有效分析，在大幅减少计算耗时的同时，还能够提供与 SRM 几乎一致的被试问题解决能力估计值。同时，综合模拟研究与实证研究的结果，我们认为 2P-ASM 比 1P-ASM 的综合表现更优；但当样本量较小(如 100 人)或任务简单(解决问题所需的操作较少)时，则推荐使用更简约的 1P-ASM。

当然，作为二分类模型，ASM 与 SRM 相比仍有一定的理论局限。比如，使用 ASM 分析行动序列数据前需要将行动序列进行二分编码，将所有错误行动序列视为“等价”，进而不可避免地损失了不同错误行动序列所提供的差异化信息。另外，由于 ASM 是对二分编码后的行动序列数据进行建模的，导致我们无法通过给定模型参数使其生成行动序列数据。

尽管本文提出两个可有效分析行动序列数据的模型，但仍有一些不足值得在今后的研究中做进一步尝试。比如，首先，与 SRM 一样，ASM 也假设被试的问题解决能力是单维的；然而，在一些问题解决任务中，有可能需要被试使用多个不同维度的问题解决能力。后续研究也可尝试进一步提出多维行动序列模型(如, 韩雨婷, 2021)。其次，在过程数据中，不仅记录了被试在问题解决各阶段所处的问题状态(即行动序列)，还记录了被试在问题解决各阶段上的时间戳信息；利用时间戳信息可以计算出被试呈现各行动序列所花费的时间，即行动时间(action times) (Fu et al., 2022)。目前，在题目层面数据分析中，已有大量关于题目作答时间(item response times)数据分析的以及将其与题目作答精度数据进行联合分析的研究(e.g., van der Linden, 2006; 2007)。后续研究也可尝试将行动时间数据与行动序列数据相结合，进一步挖掘过程数据中所包含的丰富信息(如, Fu et al., 2022)。再有，被试在解决问题过程中必须从下一个阶段的转移可选项中选择一个才能将任务继续下去；当被试不知如何选择时，是有可能通过猜测来进行选择的。后续研究也可以尝试迁移应用包含猜测参数的三参数 IRT 模型来处理行动序列数据中可能存在的猜测问题。最后，由于篇幅、时间和精力所限，模拟研究中所操纵的变量数量或水平数量有限，未能充分挖掘 ASM 在不同理想测验条件下的表现。后续研究也可尝试通过操纵其他变量(如，任务的复杂性[包含更多数量问题状态])来进一步探究 ASM 的心理计量学性能。

参考文献

Arieli-Attali, M., Ou, L., & Simmering, V. R. (2019). Understanding test takers' choices in a self-adapted test: a hidden Markov modeling of process data. *Frontiers in psychology*, *10*, 83.

Beck, L. W. (1943). The principle of parsimony in empirical science. *The Journal of Philosophy*, *40*, 617-633.

Bergner, Y., Walker, E., & Ogan, A. (2017). Dynamic Bayesian network models for peer tutoring interactions. In *Innovative assessment of collaboration* (pp. 249-268). Springer.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp.397-124). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51. https://doi.org/10.1007/BF02291411

Fu, Y., Zhan, P., Chen, Q., & Jiao, H. Joint modeling of action sequences and action times in problem-solving tasks. *PsyArXiv*. Retrieved from psyarxiv.com/e3nbc

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Stanford Univ CA Dept of Statistics*.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-511.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733-760.

Guo, J., Gabry, J., & Goodrich, B. (2018). Rjags: R interface to Stan. R package version 2.18.2. Retrieved from https://CRAN.R-project.org/package=rstan

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society: Series B (Methodological), 29*(1), 83-100.

Han, Y. (2021). The extension and application of the sequential response model for process data of problem-solving test. Unpublished doctoral dissertation, Beijing Normal University.

[韩雨婷(2021). *问题解决测验过程数据的序列反应模型的建构、拓展及其应用.* 为出版的博士论文，北京师范大学.]

Han, Y., Liu, H., & Ji, F. (2021). A sequential response model for analyzing process data on technology-based problem-solving tasks. *Multivariate Behavioral Research*, 1-18.

Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: an edit distance approach. *Journal of Educational Data Mining, 7*(1), 33-50.

Harding, S. M. E., Griffin, P. E., Awwal, N., Alom, B. M., & Scoular, C. (2017). Measuring collaborative problem solving using mathematics-based tasks. *AERA Open, 3*(3), 2332858417728046.

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education, 166*, 104170.

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In *Handbook of research on technology tools for real-world skill development* (pp. 750-777). IGI Global.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*(1), 1593-1623.

Levy, R. (2019). Dynamic Bayesian network modeling of game-based diagnostic assessments. *Multivariate*

behavioral research, 54(6), 771-794.
Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model. *Frontiers in psychology, 9,* 1372.
Liu, Y., Xu, H., Chen, Q., & Zhan, P. (2022). The measurement of problem-solving competence using process data. *Advances in Psychological Science*, *30*(3), 522-535.
[刘耀辉，徐慧颖，陈琦鹏，詹沛达. (2022). 基于过程数据的问题解决能力测量及数据分析方法. *心理科学进展, 30*(3), 522-535.]
Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-hall.
OECD (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en
Rasch, G. (1960). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Held at the Statistical Laboratory*, University of California, June 20-July 30, 1960 (Vol. 4, pp. 321). Univ of California Press.
Rosen, Y. (2017). Assessing students in human-to-agent settings to inform collaborative problem-solving learning. *Journal of Educational Measurement, 54*(1), 36-53.
Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*, 1151-1172.
Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in games*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9589.001.0001
Tang, S., Peterson, J. C., & Pardos, Z. A. (2016, April). Deep neural networks and how they apply to sequential education data. In *Proceedings of the third (2016) acm conference on learning@ scale* (pp. 321-324).
Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika, 85*(2), 378-397.
Tang, X., Wang, Z., Liu, J., & Ying, Z. (2021). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology, 74*(1), 1-33.
Van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181-204.
Van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287-308.
Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11*(12).
Wilson, M., Gochyyev, P., & Scalise, K. (2017). Modeling data from collaborative assessments: learning in digital interactive social networks. *Journal of Educational Measurement, 54*(1), 85-102.
Yuan, J., Xiao, Y., & Liu, H. (2019). Assessment of collaborative problem solving based on process stream data: A new paradigm for extracting indicators and modeling dyad data. *Frontiers in psychology, 10*, 369.
Zhan, S., Hao, J., & Davier, A. V. (2015). Analyzing process data from game/scenariobased tasks: an edit distance approach. *Journal of Educational Data Mining*, *7*(1), 33-50.
Zhan, P., & Qiao, X. (2022). Diagnostic classification analysis of problem-solving competence using process data: An item expansion method. *Psychometrika*, *87*, 1529-1547.

Zhang, S., Wang, Z., Qi, J., Liu, J., & Ying, Z. (2022). Accurate assessment via process data. *Psychometrika*, *2*, 1-22.

Zhu, M., Shu, Z., & von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment. *Journal of Educational Measurement, 53*(2), 190-211.

Binary Modeling of Action Sequences in Problem-solving Tasks: One- and Two-parameter Action Sequence Model*

Yanbin Fu[1] Qipeng Chen[1] Peida Zhan[1,2]

([1]School of Psychology, Zhejiang Normal University, Jinhua, China)

([2]Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Jinhua, China)

Abstract

Process data refers to the human-computer or human-human interaction data recorded in computerized learning and assessment systems that reflect respondents' problem-solving processes. Among the process data, action sequences are the most typical data because they reflect how respondents solve the problem step by step. However, the non-standardized format of action sequences (i.e., different data lengths for different participants) also poses difficulties for the direct application of traditional psychometric models. Han et al. (2021) proposed the SRM by combining dynamic Bayesian networks with the nominal response model (NRM) to address the shortcomings of existing methods. Similar to the NRM, the SRM uses multinomial logistic modeling, which in turn assigns different parameters to each possible action sequence in the task, leading to high model complexity. Given that action sequences in problem-solving tasks have correct and incorrect outcomes rather than equivalence relations without quantitative order, this paper proposes two action sequence models based on binary logistic modeling with relatively low model complexity: the one- and two-parameter action sequence models (1P and 2P-ASM). Unlike the SRM, which applies the NRM migration to action sequence analysis, the 1P-ASM and 2P-ASM migrate the simpler one- and two-parameter IRT models to action sequence analysis, respectively.

An illustrated example was provided to compare the performance of SRM and two ASMs with a real-world interactive assessment item, "Tickets," in the PISA 2012. The results mainly showed that: (1) the latent ability estimates of two ASMs and the SRM had high correlation; (2) ASMs took less computing time than that of SRM; (3) participants who are solving the problem correctly tend to continue to present the correct action sequences, and vice versa; and (4) compared with the fixed discrimination parameter of the SRM, the free estimated discrimination parameter of the 2P-ASM helped us to better understand the task.

A simulation study was further designed to explore the psychometric performance of the proposed model in different test scenarios. Two factors were manipulated: sample size (including 100, 200, and 500) and average problem state transition sequence length (including short and long). The SRM was used to generate the state transition sequences in the simulation study. The problem-solving task structure from the empirical study was used. The results showed that: (1) two ASMs could provide accurate parameter estimates even if they were not the data-generation model; (2) the computation time of both ASMs was lower than that of SRM, especially under

*Corresponding Author: Peida Zhan, E-mail: pdzhan@gmail.com; zhan@zjnu.edu.cn

the condition of a small sample size; (3) the problem-solving ability estimates of both ASMs were in high agreement with the problem-solving ability estimate of the SRM, and the agreement between 2P-ASM and SRM is relatively higher; and (4) the longer the problem state transition sequence, the better the recovery of problem-solving ability parameter for both ASMs and SRM.

Overall, the two ASMs proposed in this paper based on binary logistic modeling can achieve effective analysis of action sequences and provide almost identical estimates of participants' problem-solving ability to SRM while significantly reducing the computational time. Meanwhile, combining the results of simulation and empirical studies, we believe that the 2P-ASM has better overall performance than the 1P-ASM; however, the more parsimonious 1P-ASM is recommended when the sample size is small (e.g., 100 participants) or the task is simple (fewer operations are required to solve the problem).

Key words: process data, action sequence, problem state transition, action sequence model, item response theory

附录:

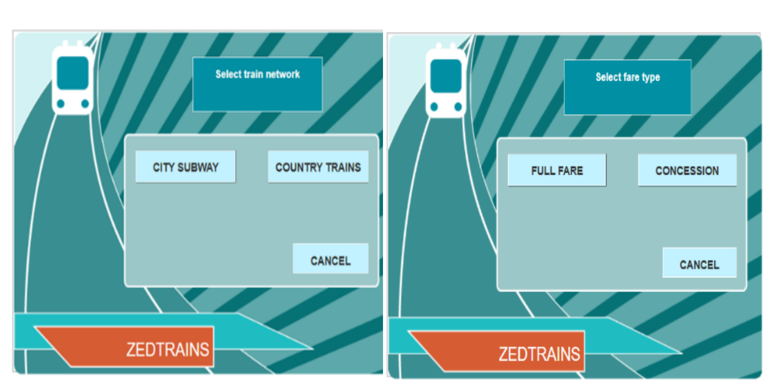
1　PISA 2012 Tickets 购票任务介绍

图 A1 是 PISA 2012 购票任务的截图，该任务包含三个子问题：

问题 1(CP038Q02)：购买一张全价的、能够乘车两次的郊区火车票，满足任务要求的被试获得 1 分，未作答或者未达成任务要求的被试得 0 分。图 A2 是完成该任务正确的行动路径。

问题 2(CP038Q01)：假设你是一位学生，可以买学生票，你计划今天在市内乘坐 4 次地铁，使用售票机找出最便宜的车票。购买了 1 张“城市地铁/优惠票/多次票/4 次”并且看过“城市地铁/优惠票/全日票”界面来比较价格的被试可以得到满分 2 分，购买了 1 张“城市地铁/优惠票/多次票/4 次”或者“城市地铁/优惠票/全日票”，但是没有同时比较过这两套方案价格的被试得 1 分，未作答或者未达到题目要求的被试得 0 分。

问题 3（CP038Q03）：假设你是一名学生，可以买学生票，你想购买一张可搭乘两次城市地铁的车票。如果被试按要求购买“城市地铁/优惠票/多次票/2 次”车票，系统会自动提示被试所选车票种类无法购买，需要更改购买方案。尝试过“城市地铁/优惠票/多次票/2 次”然后买了以下任一种车票的被试可以得 1 分：“城市地铁/全价票/多次票/1 次”，或 “城市地铁/全价票/多次票/2 次”。未作答或者未达到任务要求的被试得 0 分。



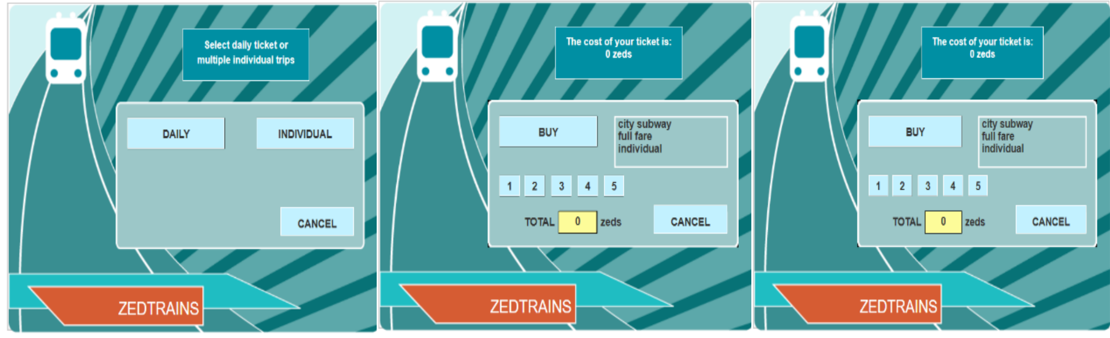
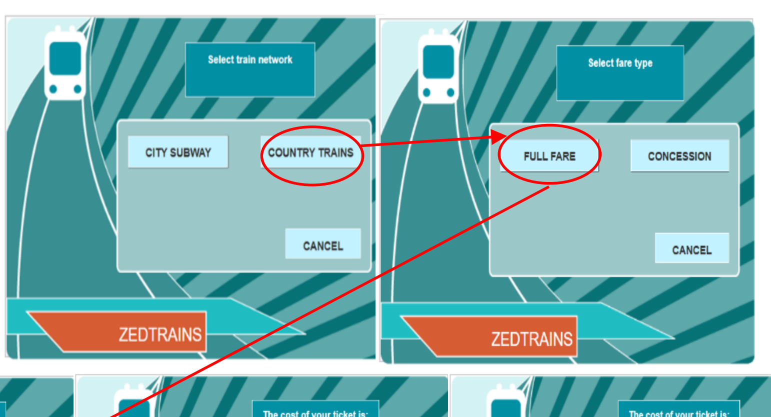


图 A1 PISA 2012 Tickets 购票任务截图



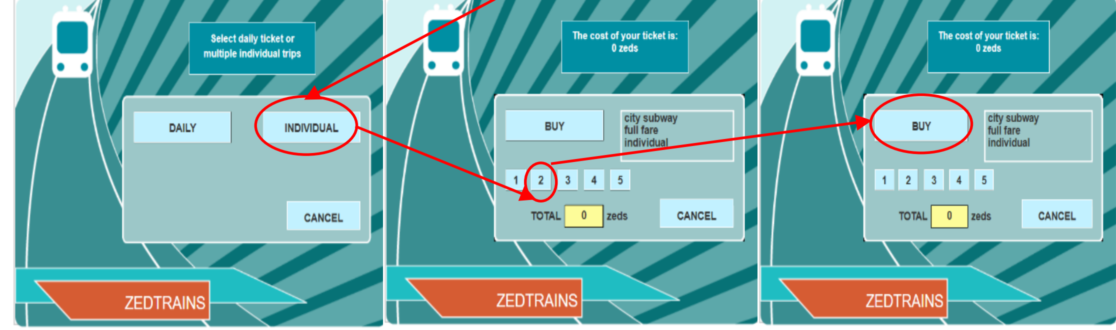


图 A2 CP038Q02 购票任务问题解决流程

注：红色箭头表示了完美解决该问题的步骤。

2 PISA 2012 购票任务数据中的异常行动序列

表 A1 展示了本文实证研究数据当中被删除的一条异常行动序列。表格中 Cnt 代表国家编号，SchoolID 代表学校编号，StdID 代表学生编号。异常的行动序列已经用红色字体标出。符合任务状态转移规则的行动序列为：Country_trains → Full_Fare → Daily → Cancel → Country_trains → Full_Fare → Individual Trip_2 → Buy。系统在记录该被试操作的过程中出错，使得被试的行动序列以倒序被重复记录了一次。限于实证研究中数据量庞大，难以对数据集中的所有行动序列一一纠正，因此不符合任务预设规则的行动序列均被删除掉了。

表 A1 异常行动序列示例

Cnt	SchoolID	StdID	Event	Time	Event Number	Action
ARE	0000068	01770	START_ITEM	843.1000	1.00	NULL
ARE	0000068	01770	ACER_EVENT	885.2000	2.00	country_trains
ARE	0000068	01770	ACER_EVENT	892.3000	3.00	full_fare
ARE	0000068	01770	ACER_EVENT	894.1000	4.00	daily
ARE	0000068	01770	ACER_EVENT	904.5000	5.00	Cancel
ARE	0000068	01770	ACER_EVENT	914.7000	6.00	country_trains
ARE	0000068	01770	ACER_EVENT	915.0000	7.00	full_fare
ARE	0000068	01770	ACER_EVENT	915.9000	8.00	individual
ARE	0000068	01770	ACER_EVENT	917.5000	9.00	trip_2
ARE	0000068	01770	ACER_EVENT	923.0000	10.00	Buy

ARE	0000068	01770	END_ITEM	928.5000	11.00	NULL
ARE	0000068	01770	END_ITEM	928.5000	12.00	NULL
ARE	0000068	01770	ACER_EVENT	923.0000	13.00	**Buy**
ARE	0000068	01770	ACER_EVENT	917.5000	14.00	**trip_2**
ARE	0000068	01770	ACER_EVENT	915.9000	15.00	**individual**
ARE	0000068	01770	ACER_EVENT	915.0000	16.00	**full_fare**
ARE	0000068	01770	ACER_EVENT	914.7000	17.00	**country_trains**
ARE	0000068	01770	ACER_EVENT	904.5000	18.00	**Cancel**
ARE	0000068	01770	ACER_EVENT	894.1000	19.00	**daily**
ARE	0000068	01770	ACER_EVENT	892.3000	20.00	**full_fare**
ARE	0000068	01770	ACER_EVENT	885.2000	21.00	**country_trains**

3 后验预测值(*ppp*)计算逻辑

对于本研究，模型拟合通过后验预测值 (*ppp*) 进行评估。选择观测值的和($O(\cdot)$)作为统计检验量，表 2 呈现了 SRM, 1P-ASM, 2P-ASM 计算该统计量的规则。值得注意的是，行动序列的观测值实际上是分类数据，我们需要比较 4,000 次 MCMC 抽样中每个样本的真值和重复抽样值，因此，行动序列的真值和抽样值将会被重新编码为 0 或 1，即 1 表示正确的状态转移，0 表示错误的状态转移，*ppp* 值即为真值的 O 统计量大于抽样值 O 统计量的比例。如果模型与数据拟合，ppp 值将接近于 0.5。

表 A2 ppp 值计算逻辑

模型		统计量
SRM	观测值	$O(S;\lambda',\theta') = \sum_{n=1}^{N}\sum_{p=1}^{P_n-1} S_{n,p} \rightarrow S_{n,p+1}$
	抽样值	$O(S^{rep\prime};\lambda',\theta') = \sum_{n=1}^{N}\sum_{p=1}^{P_n-1} S_{n,p}^{rep\prime} \rightarrow S_{n,p+1}^{rep\prime}$
1P-ASM	观测值	$O(Y;\beta',\theta') = \sum_{n=1}^{N}\sum_{p=1}^{P_n-1} Y_{n,S_p\rightarrow S_{p+1}}$
	抽样值	$O(Y^{rep\prime};\beta',\theta') = \sum_{n=1}^{N}\sum_{p=1}^{P_n-1} Y_{n,S_p\rightarrow S_{p+1}}^{rep\prime}$
2P-ASM	观测值	$O(Y;\beta',\gamma',\theta') = \sum_{n=1}^{N}\sum_{p=1}^{P_n-1} Y_{n,S_p\rightarrow S_{p+1}}$
	抽样值	$O(Y^{rep\prime};\beta',\gamma',\theta') = \sum_{n=1}^{N}\sum_{p=1}^{P_n-1} Y_{n,S_p\rightarrow S_{p+1}}^{rep\prime}$

注：λ 表示状态转移倾向参数， β 表示状态作答容易度，γ 表示状态作答区分度，θ 表示问题解决能力，$'$ 表示第 $'$ 次抽样，$S_{n,p} \rightarrow S_{n,p+1} = 1$ 或者 $Y_{n,S_p\rightarrow S_{p+1}} = 1$ 表示正确的状态转移，$S_{n,p} \rightarrow S_{n,p+1} = 0$ 或者

$Y_{n,s_p \to s_{p+1}} = 0$表示错误的状态转移。

4 实证研究中参数估计轨迹图和后验分布图

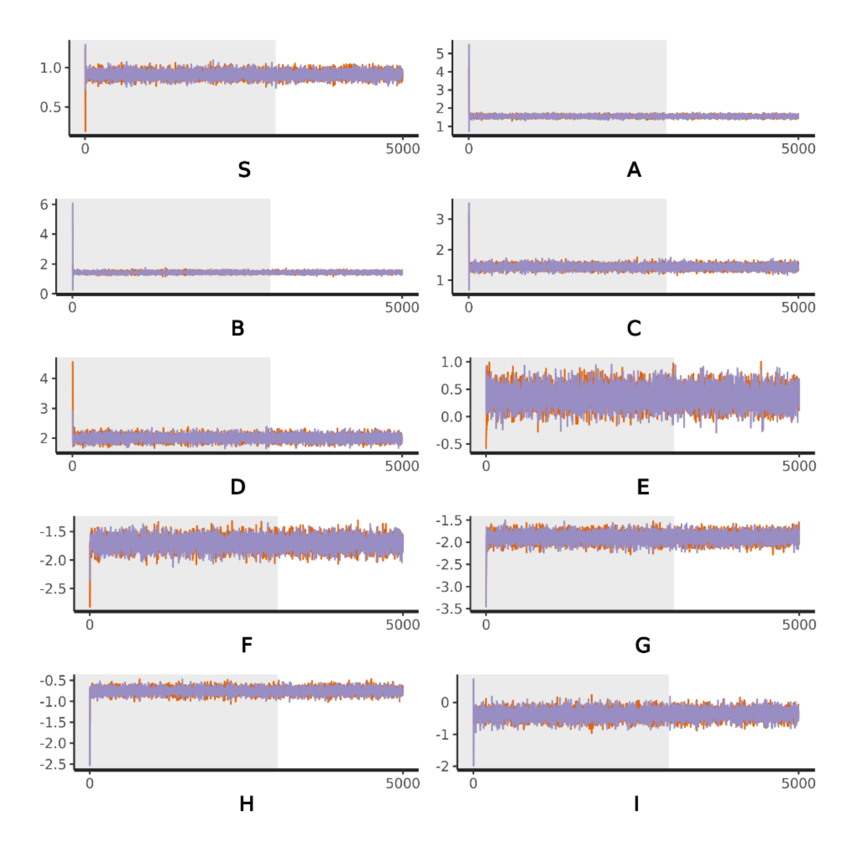


图 A3 1P-ASM 截距参数的轨迹图

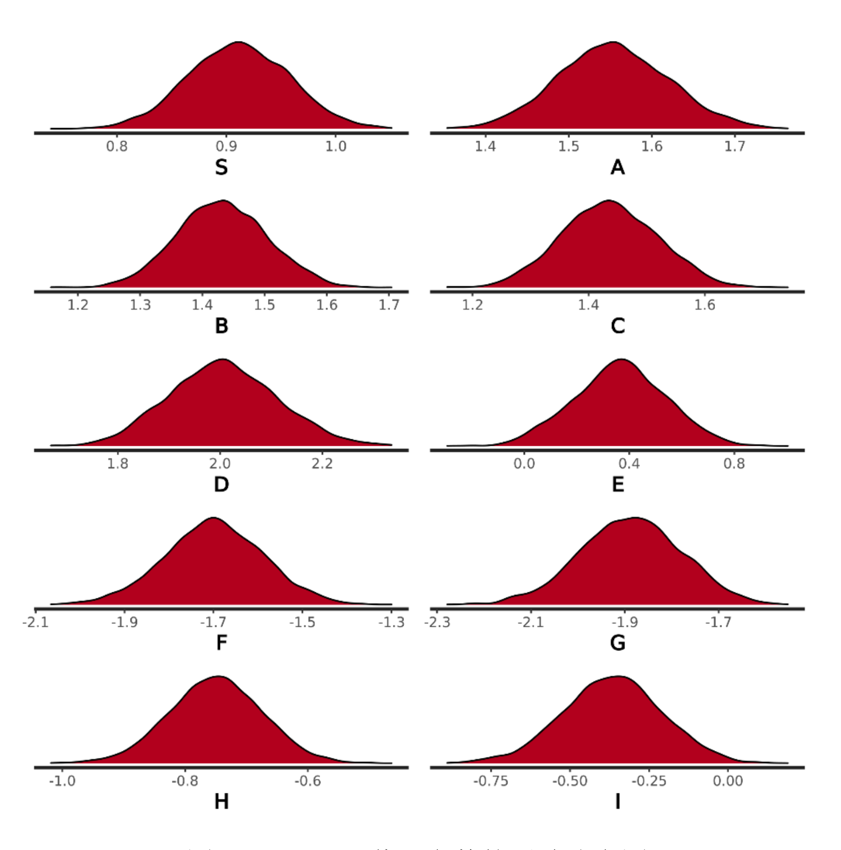


图 A4 1P-ASM 截距参数的后验分布图

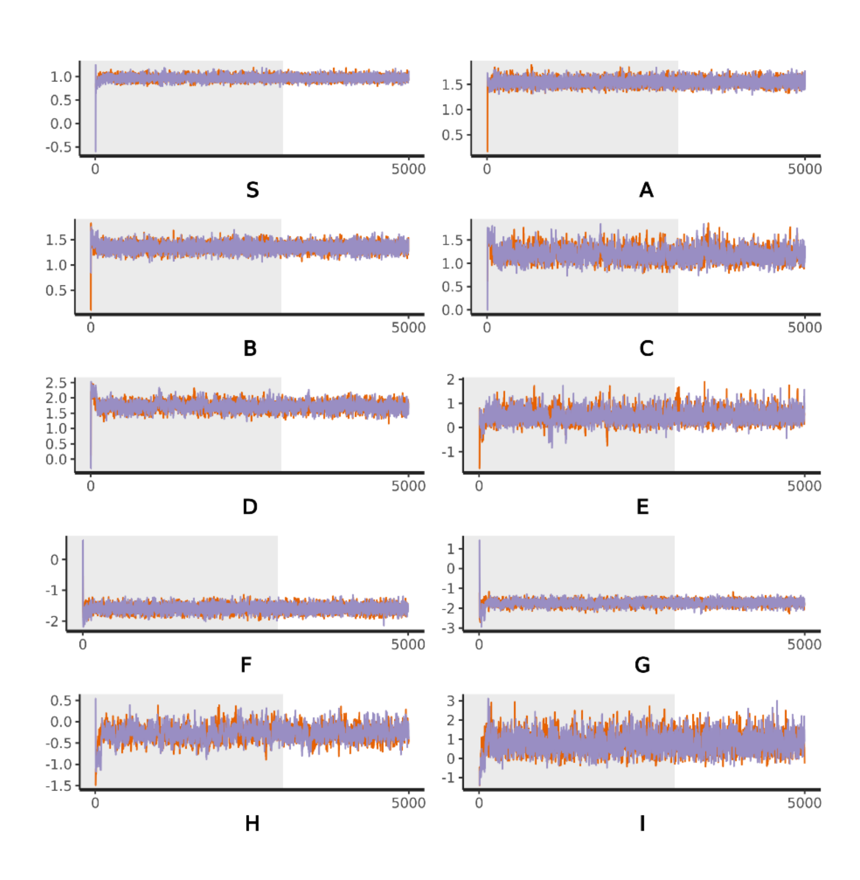


图 A5 2P-ASM 截距参数的轨迹图

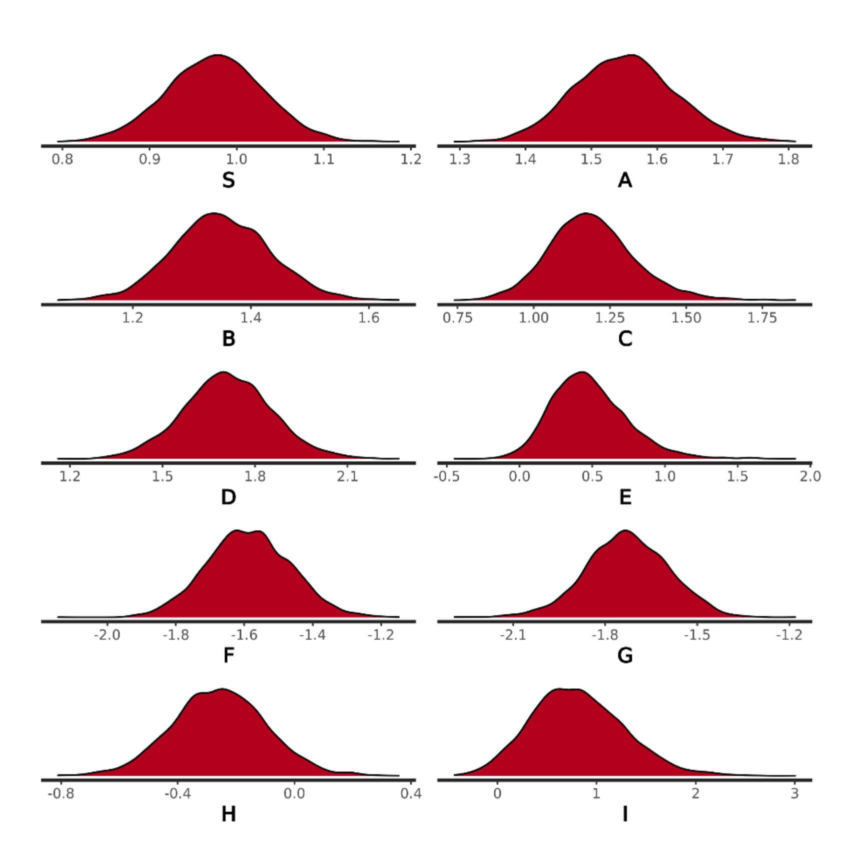


图 A6 2P-ASM 截距参数的后验分布图

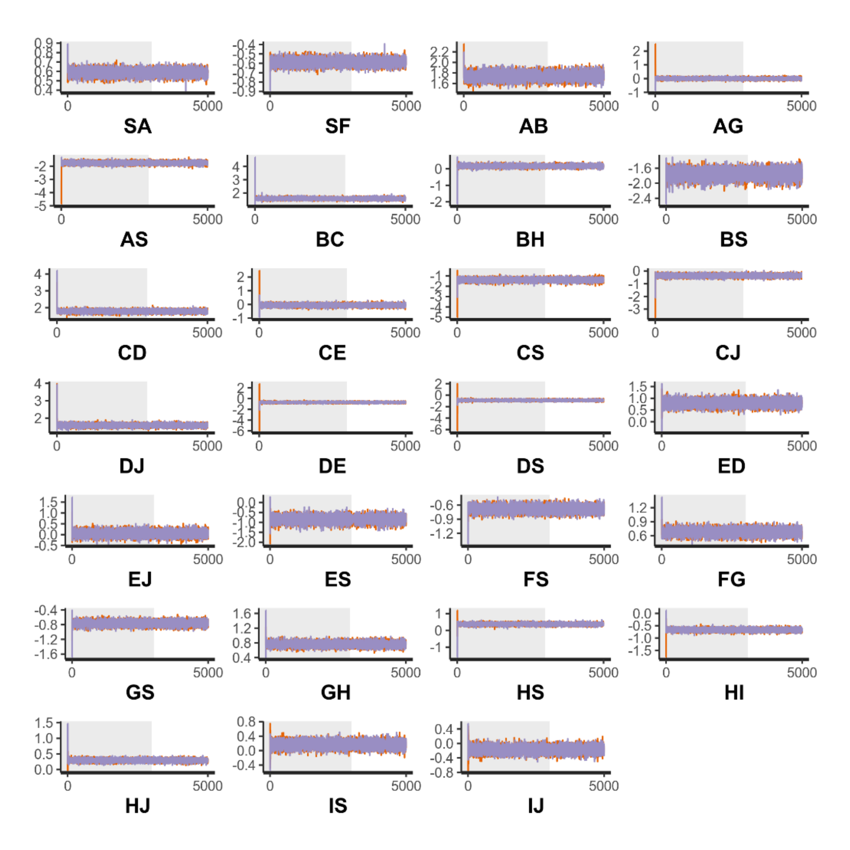


图 A7 SRM 状态转移倾向参数的轨迹图

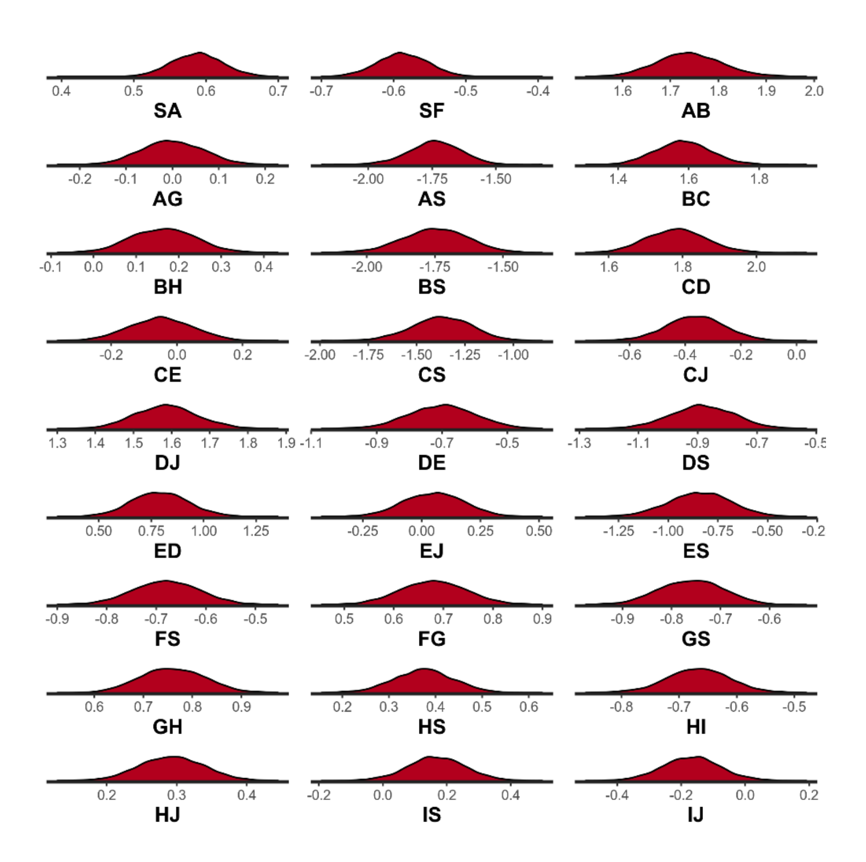


图 A8 SRM 状态转移倾向参数的后验分布图

表 A3 模拟研究中状态转移倾向参数的真值

状态转移倾向参数	短序列	长序列
$\boldsymbol{\lambda_{SA}}$	0.496	0.410
λ_{SF}	-0.469	-0.459
$\boldsymbol{\lambda_{AB}}$	1.468	1.503
λ_{AG}	-0.375	-1.096
λ_{AS}	-1.091	-0.456
$\boldsymbol{\lambda_{BC}}$	0.381	-0.932
λ_{BH}	-0.146	0.240
λ_{BS}	-0.273	0.758
$\boldsymbol{\lambda_{CD}}$	-1.001	-1.481
λ_{CE}	0.472	0.585
λ_{CS}	0.451	0.809
λ_{CJ}	0.094	0.115
$\boldsymbol{\lambda_{DJ}}$	-0.993	-1.023
λ_{DE}	0.362	0.390
λ_{DS}	0.595	0.678
$\boldsymbol{\lambda_{ED}}$	0.184	0.090
λ_{EJ}	-0.217	-0.227
λ_{ES}	0.185	0.149
$\boldsymbol{\lambda_{FS}}$	-1.001	0.965
λ_{FG}	1.013	-1.004
$\boldsymbol{\lambda_{GS}}$	-0.481	0.522
λ_{GH}	0.432	-0.599
$\boldsymbol{\lambda_{HS}}$	-0.171	0.223
λ_{HI}	0.171	-0.134
λ_{HJ}	0.028	-0.114
$\boldsymbol{\lambda_{IS}}$	-0.159	0.431
λ_{IJ}	0.071	-0.412

注：正确状态转移的参数已加粗

5 鲁棒性分析结果

表 A4 模拟研究中 1P-ASM 截距参数在不同信息水平下的估计结果

截距	有信息先验			无信息先验		
	均值	标准差	95%HPD	均值	标准差	95%HPD
β_S	0.941	0.050	(0.844, 1.038)	0.947	0.049	(0.852, 1.042)

β_A	1.613	0.068	(1.484, 1.751)	1.623	0.069	(1.492, 1.758)
β_B	-1.595	0.063	(-1.720, -1.470)	-1.603	0.062	(-1.725, -1.480)
β_C	-2.383	0.117	(-2.618, -2.161)	-2.421	0.118	(-2.650, -2.195)
β_D	-1.344	0.140	(-1.614, -1.061)	-1.377	0.143	(-1.660, -1.094)
β_E	0.069	0.168	(-0.256, 0.396)	0.070	0.170	(-0.264, 0.404)
β_F	1.801	0.087	(1.632, 1.978)	1.820	0.086	(1.654, 1.982)
β_G	0.560	0.115	(0.332, 0.782)	0.573	0.122	(0.328, 0.809)
β_H	-0.214	0.087	(-0.385, -0.040)	-0.212	0.087	(-0.385, -0.038)
β_I	0.839	0.154	(0.547, 1.139)	0.867	0.154	(0.567, 1.166)

注：有信息水平下，截距参数的先验分布为标准正态分布$N(0,1)$。无信息水平下，截距参数的先验分布服从均值为 0，标准差为 10 的正态分布$N(0,10)$。所有结果均为“500-长序列”条件下重复一次得到的估计值。

表 A5 模拟研究中 2P-ASM 截距参数在不同信息水平下的估计结果

截距	有信息先验			无信息先验		
	均值	标准差	95%HPD	均值	标准差	95%HPD
β_S	1.061	0.064	(0.937, 1.187)	1.069	0.068	(0.937, 1.210)
β_A	1.679	0.084	(1.517, 1.844)	1.694	0.086	(1.529, 1.864)
β_B	-1.910	0.109	(-2.125, -1.699)	-1.933	0.113	(-2.167, -1.723)
β_C	-2.581	0.310	(-3.220, -2.004)	-2.876	0.375	(-3.665, -2.197)
β_D	-1.026	0.305	(-1.651, -0.463)	-0.987	0.531	(-1.900, -0.005)
β_E	-0.135	0.291	(-0.719, 0.413)	-0.169	0.331	(-0.863, 0.454)
β_F	2.212	0.168	(1.904, 2.554)	2.288	0.174	(1.967, 2.643)
β_G	1.247	0.247	(0.780, 1.750)	1.374	0.268	(0.896, 1.936)
β_H	-0.139	0.104	(-0.341, 0.064)	-0.130	0.108	(-0.338, 0.078)
β_I	1.238	0.250	(0.772, 1.745)	1.356	0.275	(0.860, 1.933)

表 A6 模拟研究中 2P-ASM 斜率参数在不同信息水平下的估计结果

斜率	有信息先验			无信息先验		
	均值	标准差	95%HPD	均值	标准差	95%HPD
γ_S	2.165	0.145	(1.902, 2.463)	2.185	0.149	(1.908, 2.498)
γ_A	2.268	0.208	(1.885, 2.703)	2.308	0.213	(1.909, 2.753)
γ_B	2.075	0.186	(1.736, 2.459)	2.121	0.193	(1.760, 2.510)
γ_C	1.499	0.315	(0.921, 2.162)	1.772	0.374	(1.102, 2.560)
γ_D	0.913	0.298	(0.382, 1.537)	0.858	0.535	(0.000, 1.764)
γ_E	1.722	0.494	(0.823, 2.764)	1.810	0.562	(0.829, 3.024)
γ_F	2.125	0.246	(1.655, 2.624)	2.238	0.254	(1.770, 2.758)
γ_G	2.567	0.390	(1.866, 3.379)	2.764	0.431	(1.978, 3.695)
γ_H	2.663	0.282	(2.151, 3.254)	2.699	0.287	(2.190, 3.310)
γ_I	2.188	0.418	(1.419, 3.045)	2.383	0.466	(1.559, 3.387)

注：有信息水平下，斜率参数的先验分布服从均值为 0，标准差为 1 的对数正态分布 $\log(\gamma) \sim N(0,1)$。无信息水平下，斜率参数的先验分布服从均值为 0，标准差为 10 的对数正态分布 $\log(\gamma) \sim N(0,10)$。

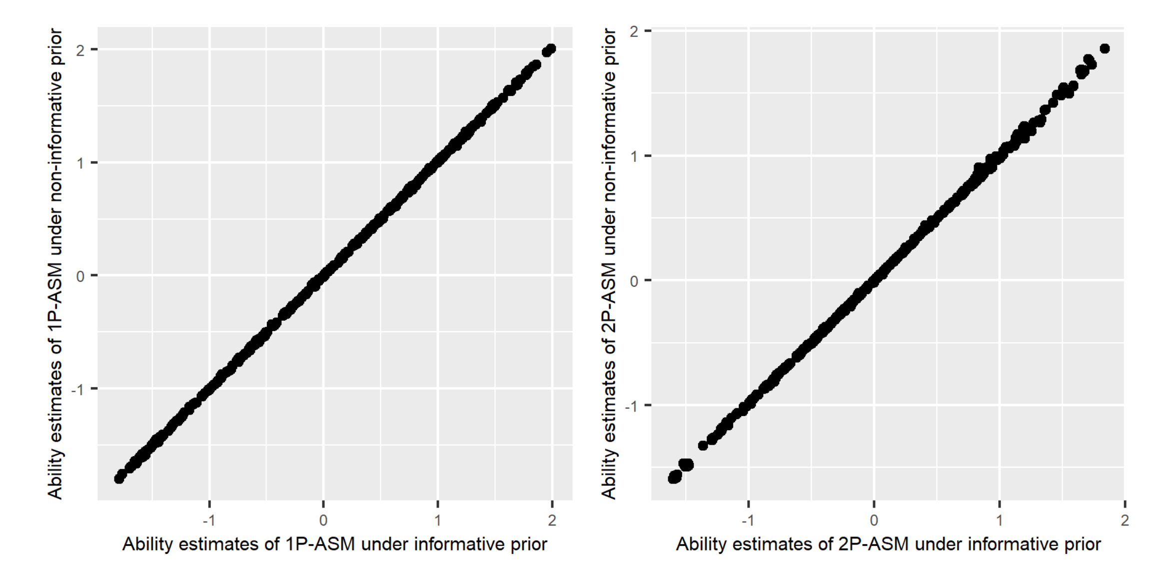


图 A9 模拟研究中 1P-ASM 和 2P-ASM 在不同信息水平下的能力估计值对比